経路探索における進路予測を利用したマルチエージェント強化学習の検討

Examination of Multi-Agent Reinforcement Learning with Trajectory Prediction in Pathfinding.

松本 天佑

Ten'yu Matsumoto

岡山大学 太田研究室

Ohta Laboratory，Okayama University

概要 本稿ではマルチエージェントシステムにおける経路探索問題を対象とし，強化学習を用いた従来手法の調査とその改良方針について議論する．具体的には，シングルエージェントによる経路探索手法である Implicit Quantile Networks（IQN）について説明し，IQN をマルチエージェントシステムに拡張した adaptive IQN の性能について報告する．また，adaptive IQN において行動のリスクを評価する際に使用する Conditional Value at Risk（CVaR）というパラメータに着目し，各エージェントの進路予測を組み込んだ新しい手法について議論する．

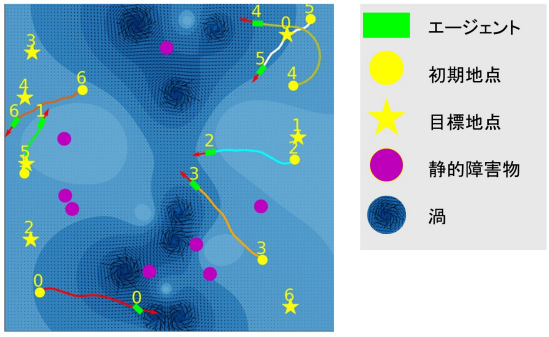

図 1: 海上の経路探索シミュレーション

1 はじめに

マルチエージェントシステム（Multi-Agent System, MAS）は，複数の独立したエージェントが相互に協力しながら目標を達成するために行動するシステムであり，その一つとしてマルチエージェント強化学習（Multi-Agent Reinforcement Learning, MARL）が存在する．MARL は複数ロボットの協調制御などの様々なタスクにおいて優れた性能を発揮する手法として知られており，例えば，自動運転車の分野などで重要な研究テーマとなっている．

本稿では，MARL を用いた経路探索手法である adaptive Implicit Quantile Networks（adaptive IQN）[6] に着目し，関連論文についての調査結果と今後の展望について議論する．

2 経路探索のシミュレーション環境

本節では，経路探索問題および adaptive IQN におけるシミュレーション環境について説明する．経路探索問題とは，特定の出発点から目的地までの最適な経路を見つけるための問題のことで，adaptive IQN では，この経路探索問題を扱う．シミュレーション環境は，海上における経路探索を再現しており，ランキン渦モデル [1] を利用した渦と，静的な障害物が存在する．これらはランダムな座標に任意の個数だけ配置される．各エージェントの初期地点と目標地点はランダムに決められ，エージェントは対応した目標地点までの経路を探索する．この際，各エージェントには障害物や目標地点の座標などの環境情報が事前に与えられておらず，周囲の一定範囲の情報を自身で取得しながら目標地点まで移動する．また，実験環境ではステアリング速度 V_S の変化によってエージェントの動作を制御する．V_S の大きさと方向の変化率を 1 ステップにわたって一定とし，エージェントのアクションを線形加速度 a と角速度 w を用いて，$a_t = (a, w)$ と表す．ただし $a \in \{-0.4, 0.0, 0.4\} m/s^2$, $w \in \{-0.52, 0.0, 0.52\}$ rad/s とする．

実験の環境のイメージを図 1に示す．図 1では，計 7 つのエージェントが，それぞれ自分に割り振られた番号に対応した目標地点までの経路を探索する．

3 Implicit Quantile Networks

Implicit Quantile Networks（IQN）は，シングルエージェントによる経路探索手法であり，Deep Q-Network（DQN）[3] をはじめとする従来手法が行動に対する報酬の期待値を学習するのに対し，行動に対する獲得報酬量を確率分布で表現して学習するという特徴がある [4]．

IQN ではニューラルネットワーク（NN）で学習することによって，行動に対する獲得報酬量の確率分布を近似する．NN の学習では，NN を利用して予測した報酬値と，実際の報酬値を基に損失を求め，誤差逆伝播法を利用して損失が最小となるように NN のパラメータを調整する．

また，IQN において損失は以下の式で計算される．

$$\mathcal{L}_{\mathrm{IQN}} = \frac{1}{N'} \sum_{i=1}^{N} \sum_{j=1}^{N'} \rho_{\tau_i}^{\kappa} \left(\delta^{\tau_i, \tau_j'} \right). \tag{1}$$

$\delta^{\tau_i, \tau_j'}$ は Temporal Difference Error, $\rho_{\tau_i}^{\kappa}$ は Quantile Huber Loss である．Temporal Difference Error は予

想報酬と実際にもらえた報酬の差を表す数値であり，Quantile Huber Loss は，以下の式で表される関数である．

$$\rho_\tau^\kappa(u) = \left|\tau - \mathbf{1}_{\{u<0\}}\right| \left(\mathcal{L}_\kappa(u)/\kappa\right), \\ \text{where } \mathcal{L}_\kappa(u) = \begin{cases} \frac{1}{2}u^2, & \text{if } |u| \le \kappa \\ \kappa\left(|u| - \frac{1}{2}\kappa\right), & \text{otherwise} \end{cases}. \quad (2)$$

τ は分位数であり，0 から 1 の任意の数値を自分で設定する．κ は閾値であり任意の数値を自分で設定する．閾値以下の誤差には二乗誤差を用い、それを上回る誤差に対しては絶対誤差を用いる．N は，将来の報酬の予想のサンプル数，N' は実際に貰った報酬のサンプル数を表している．また，$\mathbf{1}_{\{u<0\}}$ は指示関数であり $u < 0$ の時に 1，それ以外のときに 0 を取る．

また IQN は損失を算出する際に使用するパラメータとして CVaR threshold value（CVaR 閾値）を導入した．CVaR 閾値は ϕ で表され，$0 \le \phi \le 1$ の任意の実数を取る．式 (1) で表したように損失の算出には，NN で予測された報酬の分布と実際の報酬の分布から取り出した幾つかのサンプルを使用するが，CVaR 閾値を利用する場合，使用するサンプルを報酬分布における ϕ 分位点以下のもののみとする．これにより，CVaR 閾値を下げた場合は，より報酬の少ない一部のサンプルのみを利用して学習を行うため，少ない報酬における報酬評価の精度が向上することが期待される．

図 2に異なる VaR 閾値で利用される報酬のサンプルの例を示す．これらの図はある状態で特定の行動を取った際に得られる報酬とその確率をグラフで表したものであり，損失の学習には，この中からランダムに幾つかのサンプルを抽出するものとする．CVaR 閾値を赤色の点線と青色の点線で表しており，サンプルは ϕ 分位点以下のものから抽出する．赤色の点線の CVaR 閾値は 1.0, 青色の点線の CVaR 閾値は 0.25 である．このように任意の CVaR 閾値を設定することで，学習を行う際に報酬の少ない部分に着目するか報酬の多い部分に着目するかを決めることができる．

4 adaptive IQN

4.1 メカニズムの概要

Lin らは，訓練エージェント間のパラメータ共有によって IQN を MAS に拡張し，適切な経路を探索する adaptive IQN を提案した [5, 6]．adaptive IQN は，IQN における CVaR 閾値を動的に変化させることで，マルチエージェントシステムにおける協調制御を実現させている．CVaR 閾値を ϕ とすると ϕ は以下の式で計算される．

$$\phi = \begin{cases} \frac{\min(d(X,\ XO))}{d_0} & \text{if } \min(d(X,\ XO)) \le d_0 \\ 1.0 & \text{if } \min(d(X,\ XO)) > d_0 \end{cases}. \quad (3)$$

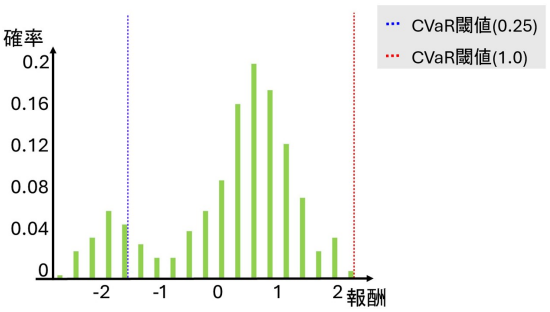

図 2: CVaR 閾値と選択サンプルのイメージ

ここで X は自身，XO はそれ以外のエージェントや障害物の座標であり，$\min(d(X,\ XO))$ は，X と XO の最小距離，つまり自分と最も近い物体との距離を表す．式 (3) をもとに CVaR 閾値を動的に変更することにより，周囲に他のエージェントや障害物が存在しない場合はより報酬の期待値が高い行動を取り，障害物などが近づくにつれ，よりリスクの低い行動を取るように動的に行動を変化させることができる．

4.2 他手法との比較実験

adaptive IQN の有効性を検証するため, IQN, DQN, Artificial Potential Field（APF）[7] , Reciprocal Velocity Obstacle（RVO）[2] の計 4 つの強化学習と精度を比較した [5, 6]．評価指標は探索成功率，探索成功した場合の探索終了までの平均時間，探索成功した場合の消費エネルギーである．消費エネルギーは，プロセス中に実行された全行動の大きさを合計して求める．

実験の結果，adaptive IQN は探索成功率において他の全ての手法を上回り，平均時間と消費エネルギーに関しても優れた結果を示した．特に探索成功率では，実験環境の複雑さに応じて比較手法が著しく減少するのに対し，adaptive IQN の減少は緩やかであった．

5 まとめと今後の課題

本稿では，シングルエージェントによる経路探索手法である IQN について説明をした上で海上の経路探索問題についてマルチエージェント強化学習 adaptive IQN の論文を調査した．adaptive IQN は以下の特徴を持つ．

- CVaR 閾値を用いて学習に使うサンプルの範囲を限定する．
- CVaR 閾値を，周囲の物体との距離に応じて変動させる．

また，その課題として adaptive IQN は CVaR 閾値の算出に周囲の物体との距離のみを利用していたが，より多彩な情報を利用することで，CVaR 閾値の数値をさらに適切に設定できると考えられる．

例えば使用する情報の具体的な例として周囲の物体との相対速度が挙げられる．つまり周囲の物体との相

対速度と距離を求め進路を予測することで，このまま移動を続けた場合の衝突可能性または衝突までの時間が求められるため，現在の衝突のリスクをより正確に評価し，探索の精度を高めることが期待できる．

今後の課題として，物体同士の相対速度を基にした進路予測を組み込むように CVaR 閾値のメカニズムを改良し，計算される衝突リスクに基づくマルチエージェント強化学習を提案することを考えている．

参考文献

[1] David John Acheson, *Elementary Fluid Dynamics*. Acoustical Society of America, 1991.

[2] Jur van den Berg *et al.*, “Reciprocal velocity obstacles for real-time multi-agent navigation,” *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1928–1935, 2008

[3] Volodymyr Mnih *et al.*, “Human-level control through deep reinforcement learning,” *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[4] Will Dabney *et al.*, “Implicit Quantile Networks for Distributional Reinforcement Learning,” *Proc. of International Conference on Machine Learning 2018 (ICML 2018)*, 2018.

[5] Xi Lin *et al.*, “Robust Unmanned Surface Vehicle Navigation with Distributional Reinforcement Learning,” *Proc. of The 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2023)*, 2023.

[6] Xi Lin *et al.*, “Decentralized Multi-Robot Navigation for Autonomous Surface Vehicles with Distributional Reinforcement Learning,” *Proc. of The 2024 IEEE International Conference on Robotics and Automation (ICRA 2024)*, 2024.

[7] Xiaojing Fan *et al.*, “Improved artificial potential field method applied for AUV path planning,” *Mathematical Problems in Engineering*, vol. 2020, pp. 1–21, 2020.